Homology Induction: the use of machine learning to improve sequence similarity searches

Karwath, Andreas; King, Ross Donald

Published in:
BMC Bioinformatics

DOI:
10.1186/1471-2105-3-11

Publication date:
2002

Citation for published version (APA):
Karwath, A., & King, R. D. (2002). Homology Induction: the use of machine learning to improve sequence similarity searches. *BMC Bioinformatics*, *3*(11). https://doi.org/10.1186/1471-2105-3-11

species(Name, Species).

Describes the organisms species declaration in one string, e.g. species(p02102, capra_hircus_goat).

classification(Name, Classification_class).

This declaration defines the peptides' organisms origin classification. The classification is divided from top to bottom of the phylogenic tree, starting with the most general classification and ending with the most specific classification. To improve computational efficiency, each leaf of the tree is an entry in the database. The following example classifications are taken from the SWISS-PROT protein P02102.

- classification(p02102, eukaryota).
- classification(p02102, metazoa).
- classification(p02102, chordata).
- classification(p02102, vertebrata).
- classification(p02102, tetrapoda).
- classification(p02102, mammalia).
- classification(p02102, eutheria).
- classification(p02102, artiodactyla).

desc(Name, Description).

Defines the protein's description word by word.

keyword(Name, Keyword).

Defines the proteins keyword.

mol_wt_rule(Name, Mol_weight_interval).

The relative molecular weight of the protein.

db_ref(Name, Database_identyfier, Primary_identyfier,Secondary_identyfier,Status).

Defines a database reference to a specific database and its entry.

domain_rule(Name, Query_start_interval, Query_end_interval, Target_start_interval, Target_end_interval).

Defines the relative hit taken from the local PSI-BLAST alignment.

amino_acid_ratio_rule(Name, Amino_acid, Percentage_interval).

Defines the content of a particular amino acid on a 1-10 scale.

amino_acid_pair_ratio_rule(Name, Amino_acid, Amino_acid, Promille_interval).

Measures the pairwise amino acid content on a 1-10 scale.

seq_length_rule(Name, Seq_length_interval).

Measures the total number of amino acids in the sequence on a scale from 1 to 10.

signalip1_rule(Name, Signalip1_interval).

Defines the predicted cleavage sites found using the SignalIP program (Nielsen et. al, 1997) server. Signalip1 is the protein's maximum cleavage site score.

signalip2_rule(Name, Signalip2_interval).

Signalip2 is the protein's maximum combined cleavage site score

signalip3_rule(Name,Signalip3_interval).

Signalip3 is the protein's maximum signal peptide score

signalip4_rule(Name, Signalip4_interval1, Signalip4_interval2).

The signalip4 interval is the proteins most likely cleavage site

hydro_cons_rule(Name, hc_interval1, hc_interval2, hc_interval3, hc_interval4).

hc_interval1 defines the average hydrophobic moment assuming α-helix,
hc_interval2 the average hydrophobic moment assuming β-strand, and
hc_interval3 the average hydropathy using the Kyte-Doolittle hydrophilicity scale

sec_struc_rule(Name, Sec_position_interval, Secondary_structure, Sec_length_interval).

Defines the position, length, and type of predicted secondary structure.

sec_struc_alpha_rule(Name, Sec_position_alpha_interval,Sec_length_alpha_interval).

Defines the position and length of predicted secondary structure of type α.

sec_struc_beta_rule(Name, Sec_position_beta_interval, Sec_length_beta_interval).

Defines the position and length of predicted secondary structure of type β.

sec_struc_coil_rule(Name, sec_position_coil_interval, sec_length_coil_interval).

Defines the position and length of predicted secondary structure of type coil.

sec_struc_distribution_rule(Name,Secondary_structure,Sec_dist_interval).

Measures the distribution of all three types of secondary structure predictions

sec_struc_conf_rule(Name,Sec_conf_interval).

Measures the confidence in the over all secondary structure prediction.

sec_struc_conf_alpha_rule(Name,Sec_conf_alpha_interval).

Measures the confidence in the α secondary structure predictions.

sec_struc_conf_beta_rule(Name,Sec_conf_beta_interval).

Measures the confidence in the β secondary structure predictions.

sec_struc_conf_coil_rule(Name,Sec_conf_coil_interval).

Measures the confidence in the coil secondary structure predictions.

Table 1. The definition of types of data held in the logical database. The data was obtained from a wide variety of bioinformatic sources. This information was selected for relevance to the detection of homology. For each protein we collected information for each type of data if it was available.